Multi-Classification by Categorical Features via Clustering

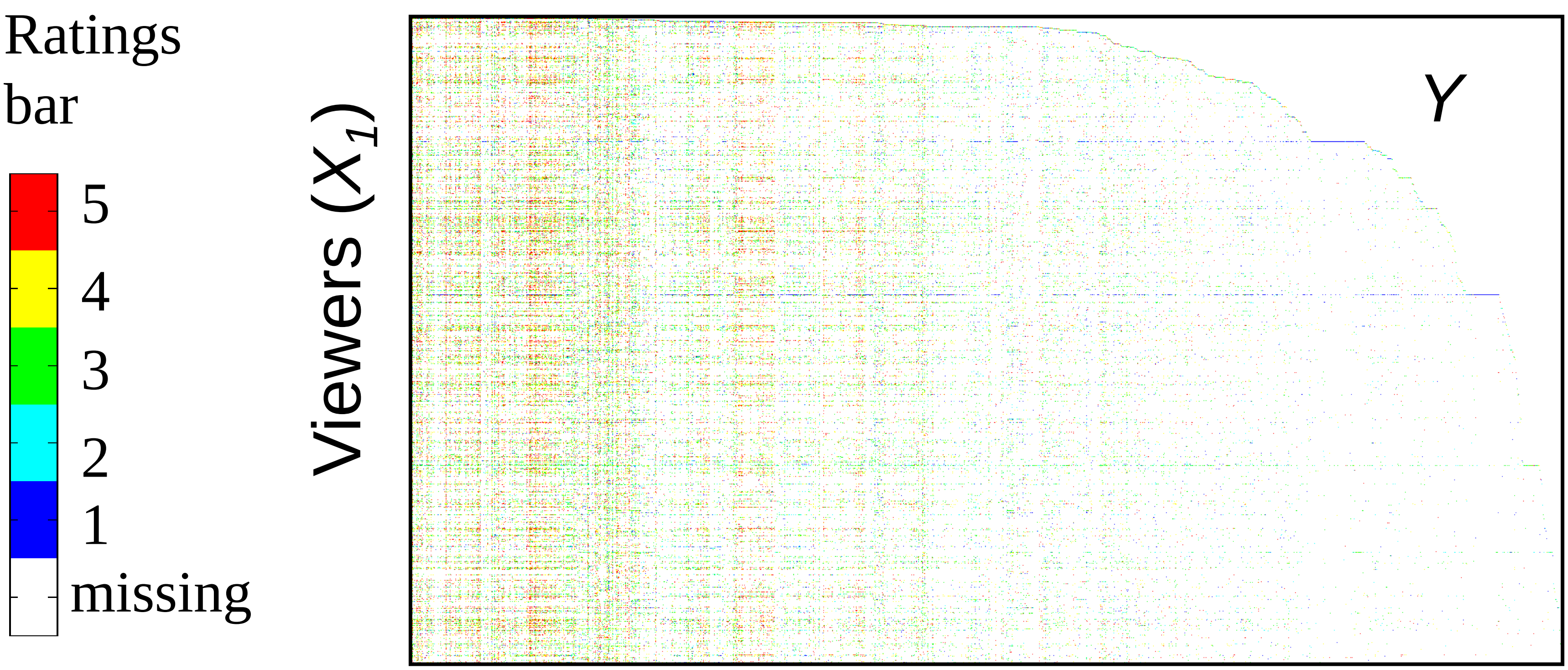
Yevgeny Seldin

Joint work with Naftali Tishby

The Hebrew University of Jerusalem

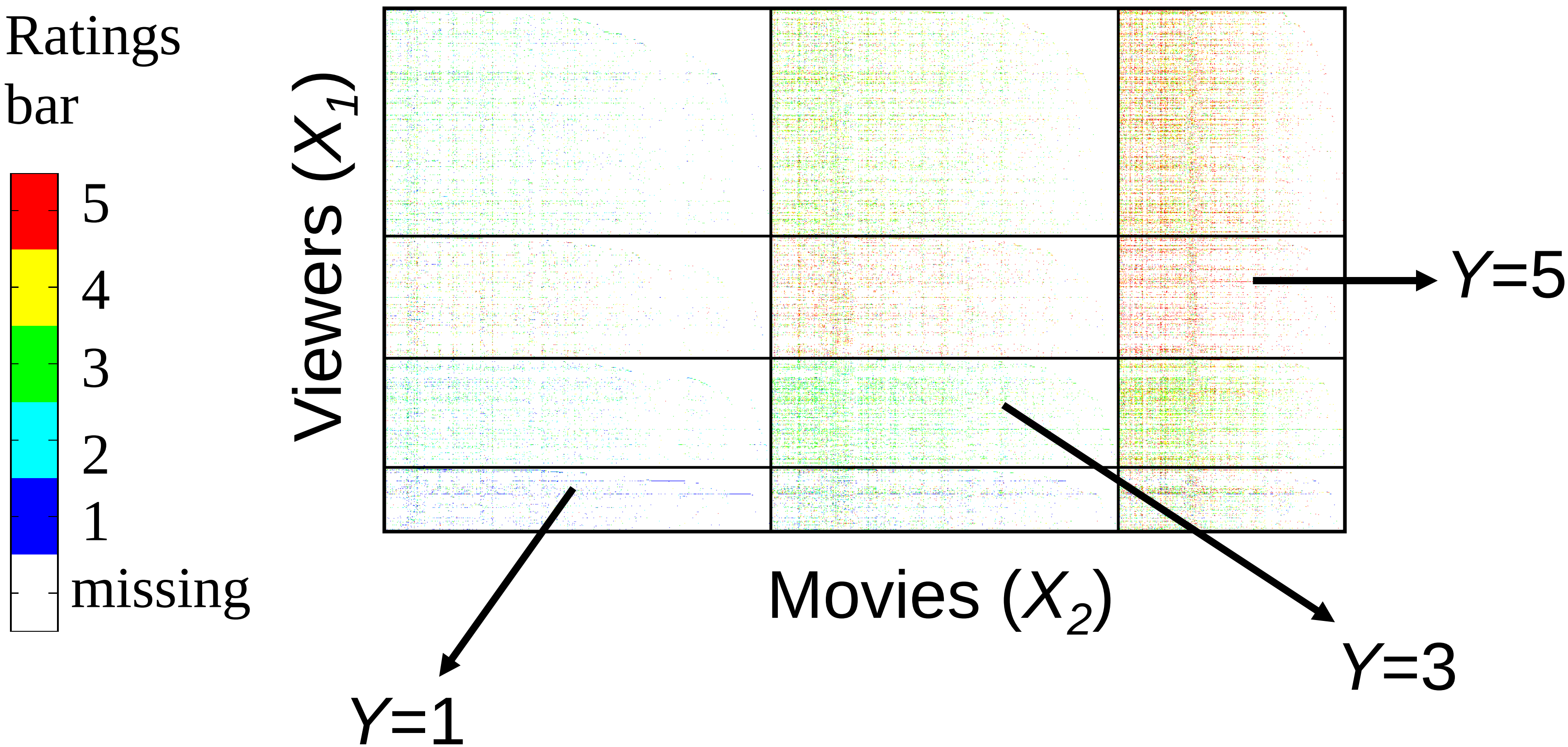
Multi-Classification
by Categorical Features

Example: Collaborative Filtering



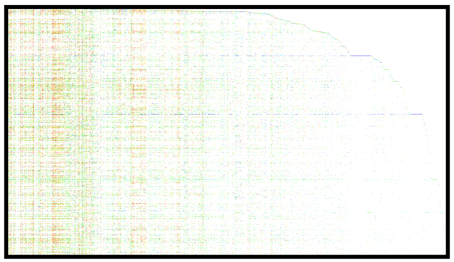
MovieLens Data

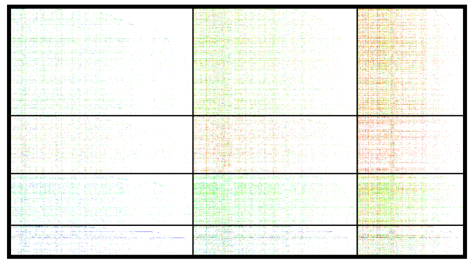
Multi-Classification via Grid Clustering

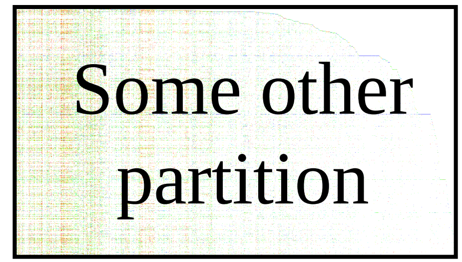


[Seldin, Slonim & Tishby, NIPS06]

Question: which partition is better?







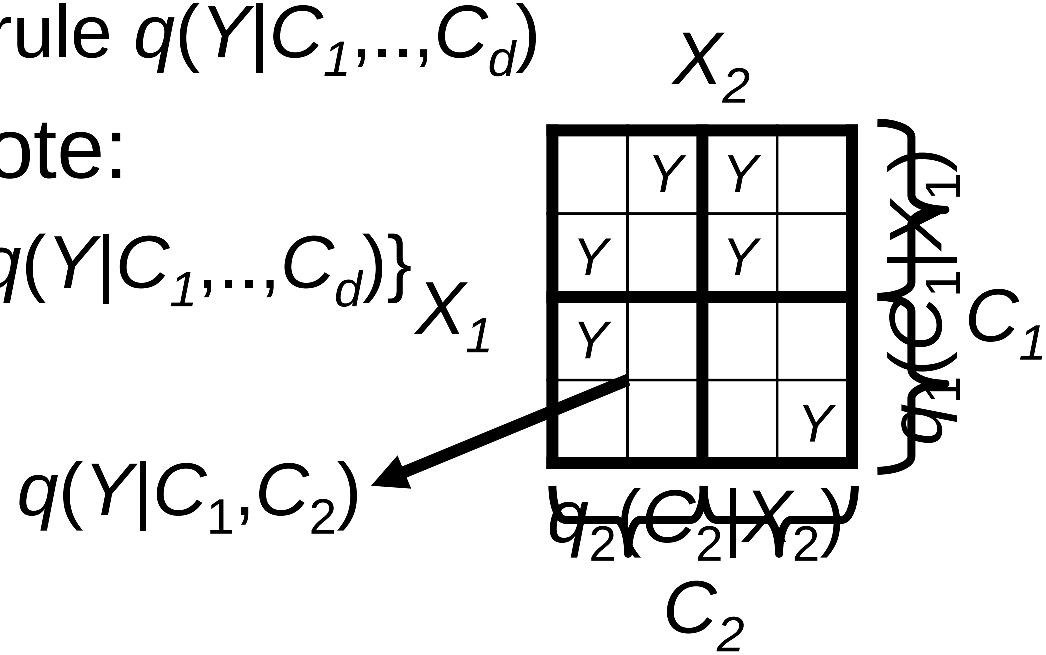
- Statistical Reliability vs. Precision Tradeoff

The Approach We Take

- Relate with generalization properties

Some Definitions

- Classification via Stochastic Grid Clustering:
  - A set of stochastic mappings $q_i(C_i|X_i)$
  - A classification rule $q(Y|C_1,..,C_d)$
- Collectively denote:
  - $Q = \{\{q_i(C_i|X_i)\}, q(Y|C_1,..,C_d)\}$



Generalization Bound

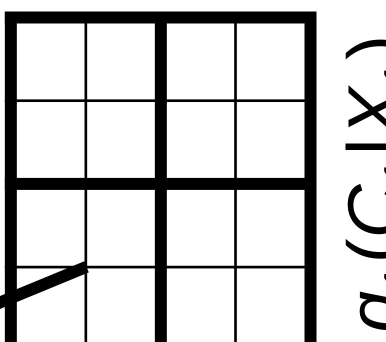
- With probability ≥ 1-δ:

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{\sum_i n_i I_U(X_i;C_i) + K}{2N}}$$

Expected loss of Q

Empirical loss of Q

Complexity term



$q(Y|C_1,C_2)$

$q_2(C_2|X_2)$

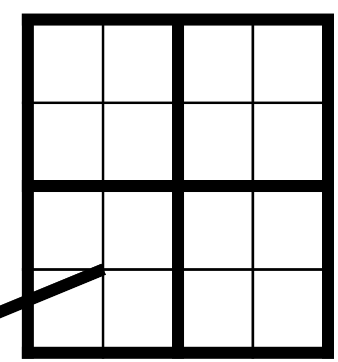
$q_1(C_1|X_1)$

Generalization Bound

- With probability ≥ 1-δ:

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{\sum_i n_i I_U(X_i;C_i) + K}{2N}}$$

$n_i=|X_i|$



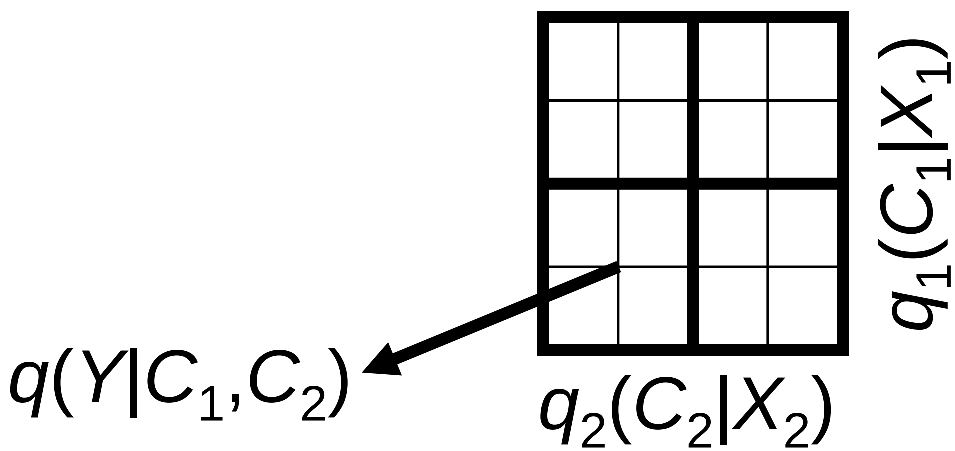
Generalization Bound

- With probability ≥ 1-δ:

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{\sum_i n_i I_U(X_i;C_i) + K}{2N}}$$

Information preserved by $q_i(C_i|X_i)$

(w.r.t. $p(X_i)=1/n_i$)

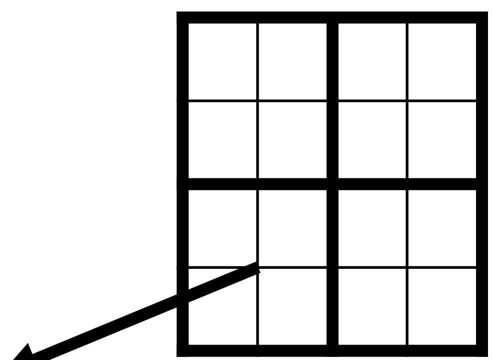


Generalization Bound

- With probability ≥ 1-δ:

$$L(Q) \le \hat{L}(Q) + \sqrt{\frac{\sum_i n_i I_U(X_i;C_i) + K}{2N}}$$

Independent of $q_i(C_i|X_i)$

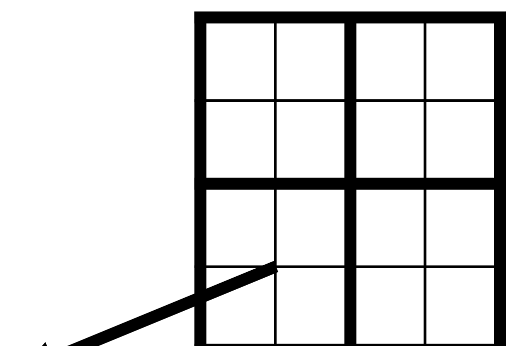


Generalization Bound

- With probability ≥ 1-δ:

$$L(Q) \le \hat{L}(Q) + \sqrt{\frac{\sum_i n_i I_U(X_i;C_i) + K}{2N}}$$

Sample size



$q(Y|C_1,C_2)$

$q_2(C_2|X_2)$

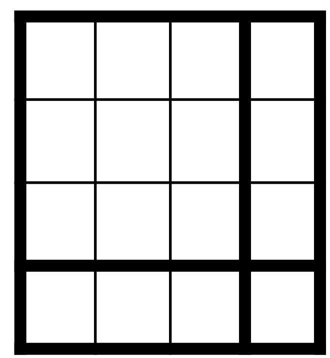
$q_1(C_1|X_1)$

Generalization Bound

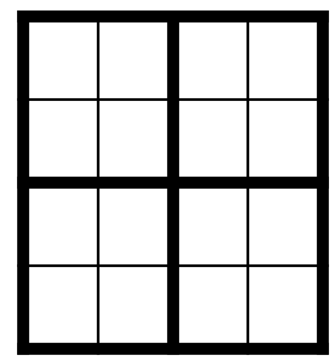
- With probability ≥ 1-δ:

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{\sum_i n_i I_U(X_i;C_i) + K}{2N}}$$

<u>Optimization tradeoff:</u>
Empirical loss vs.
"Effective" partition
complexity



Lower Complexity



Higher Complexity

Generalization Bound

- With probability ≥ 1-δ:

$$L(Q) \le \hat{L}(Q) + \sqrt{\frac{\sum_i n_i I_U(X_i;C_i) + K}{2N}}$$

$$K = \underbrace{\sum_i \left(m_i \ln n_i + \frac{(\ln n_i + 1)^2}{4} \right)}_{\text{Logarithmic in } n_i} + \underbrace{\left(\prod_i m_i \right)}_{\text{Number of partition cells}} \ln |Y| + \underbrace{\frac{1}{2} \ln(4N) + \ln \frac{1}{\delta}}_{\text{“usual stuff”}}$$

$m_i = |C_i|$

Proof Idea

- Start with the PAC-Bayesian bound:

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{D(Q \| P) + \ln(4N)/2 + \ln(1/\delta)}{2N}}$$

  – [McAllester 99], [Maurer 04]

Proof Idea

- Start with the PAC-Bayesian bound:

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{D(Q \| P) + \ln(4N)/2 + \ln(1/\delta)}{2N}}$$

  - [McAllester 99], [Maurer 04]

- Design a combinatorial prior *P(h)* by counting the number of hard partitions

Proof Idea

- Start with the PAC-Bayesian bound:

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{D(Q \| P) + \ln(4N)/2 + \ln(1/\delta)}{2N}}$$

  – [McAllester 99], [Maurer 04]

- Design a combinatorial prior *P(h)* by counting the number of hard partitions
- Calculate *D(Q||P)*

Proof Idea

- Start with the PAC-Bayesian bound:

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{D(Q \| P) + \ln(4N)/2 + \ln(1/\delta)}{2N}}$$

  – [McAllester 99], [Maurer 04]
- Design a combinatorial prior *P(h)* by counting the number of hard partitions
- Calculate *D(Q||P)*
- Details: at the paper/poster

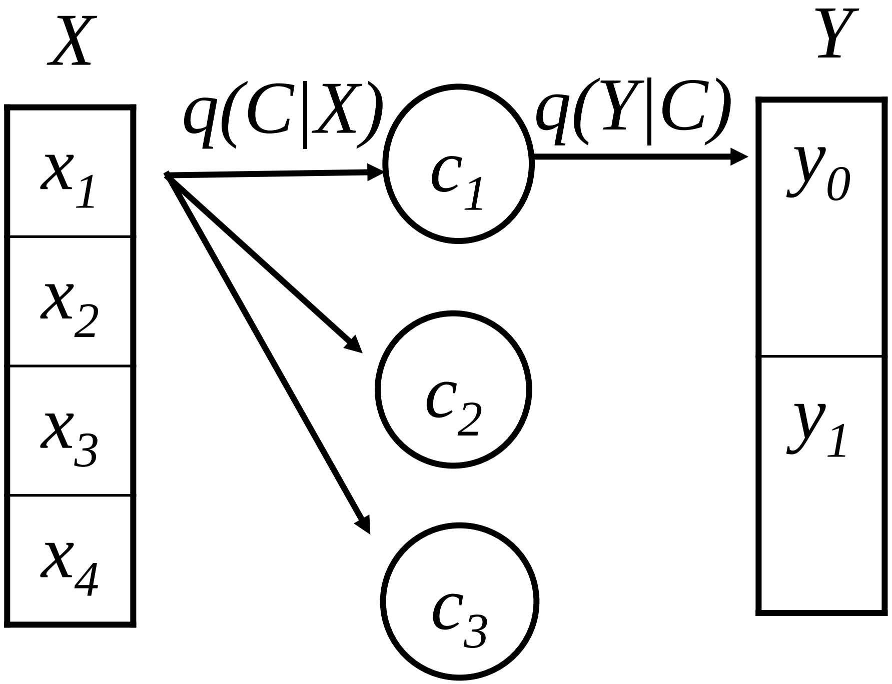
Messages

- For Clustering:
  - Evaluate clustering by its generalization properties on the task it is designed for
- For Classification:
  - Unify feature values to amplify statistical reliability

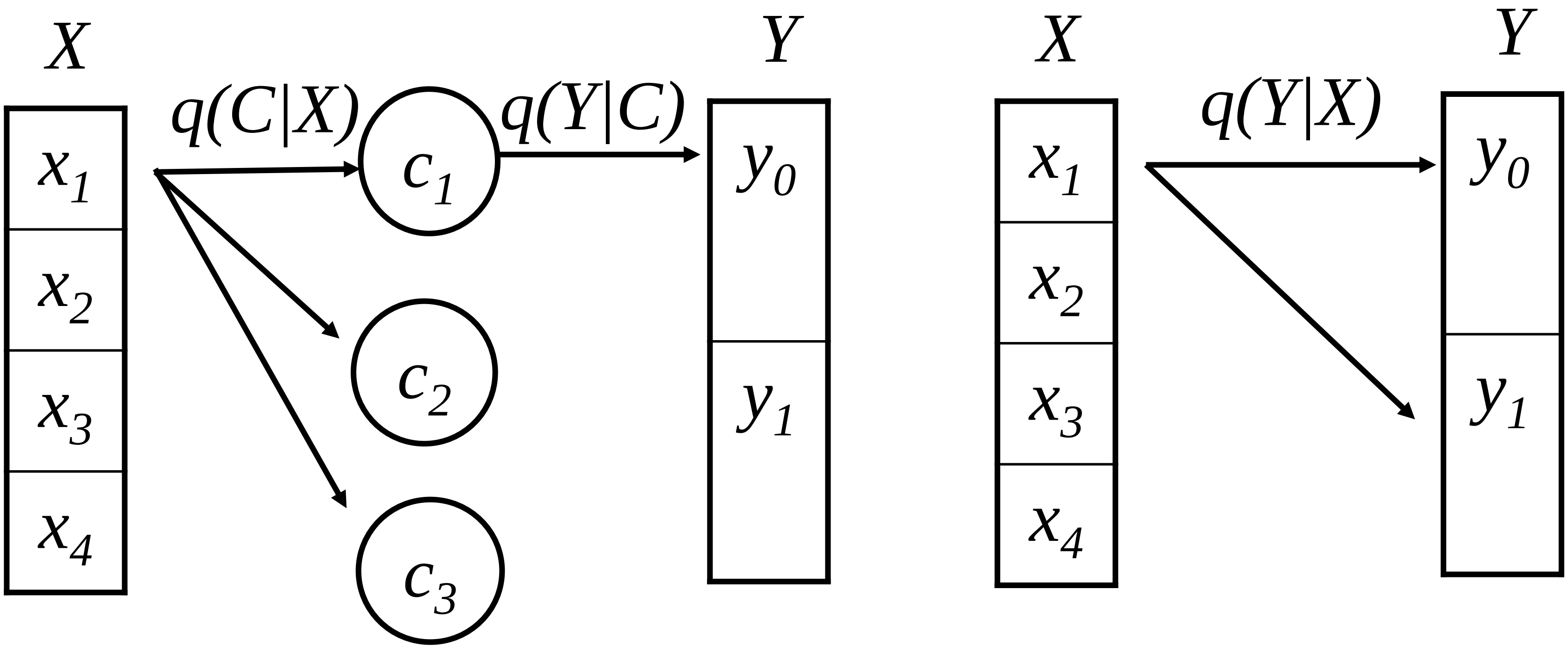
Classification by a Single Feature

- A tighter and simpler bound
- Application: Feature Ranking

Classification by a Single Feature

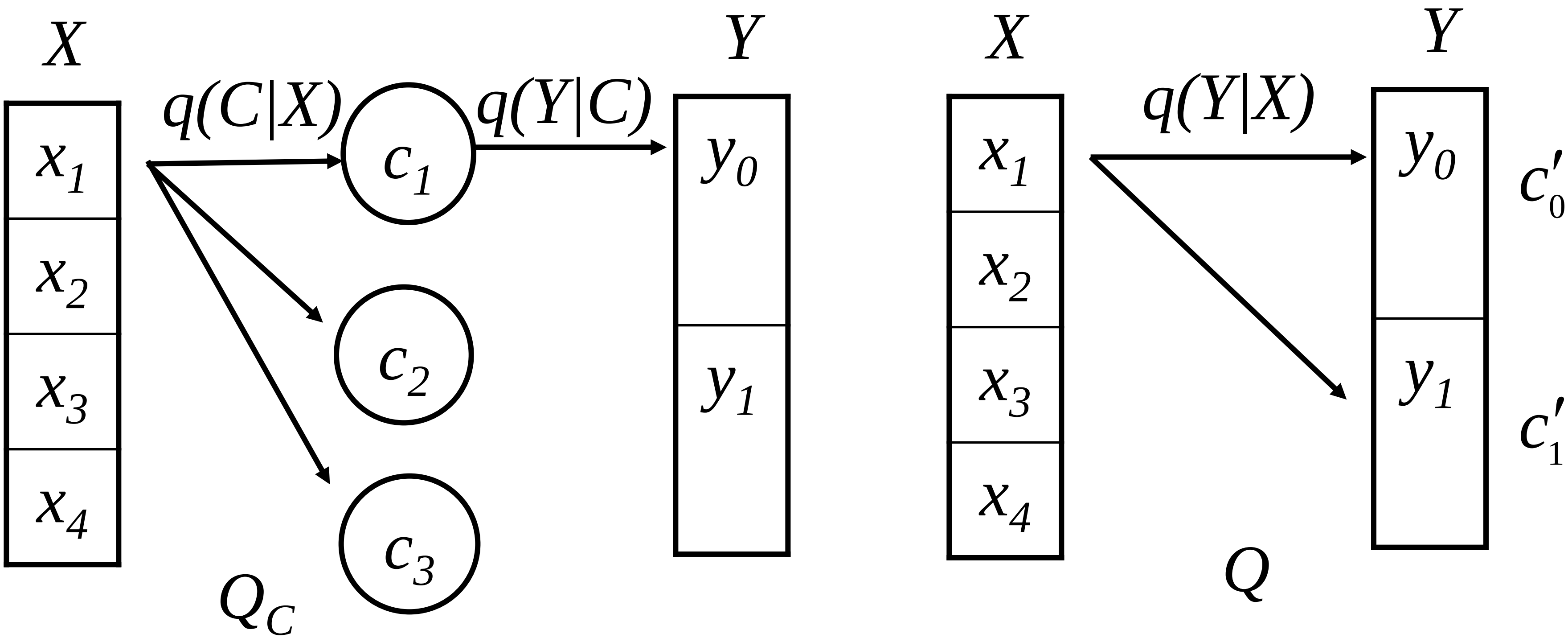


Equivalent “Direct Mapping”



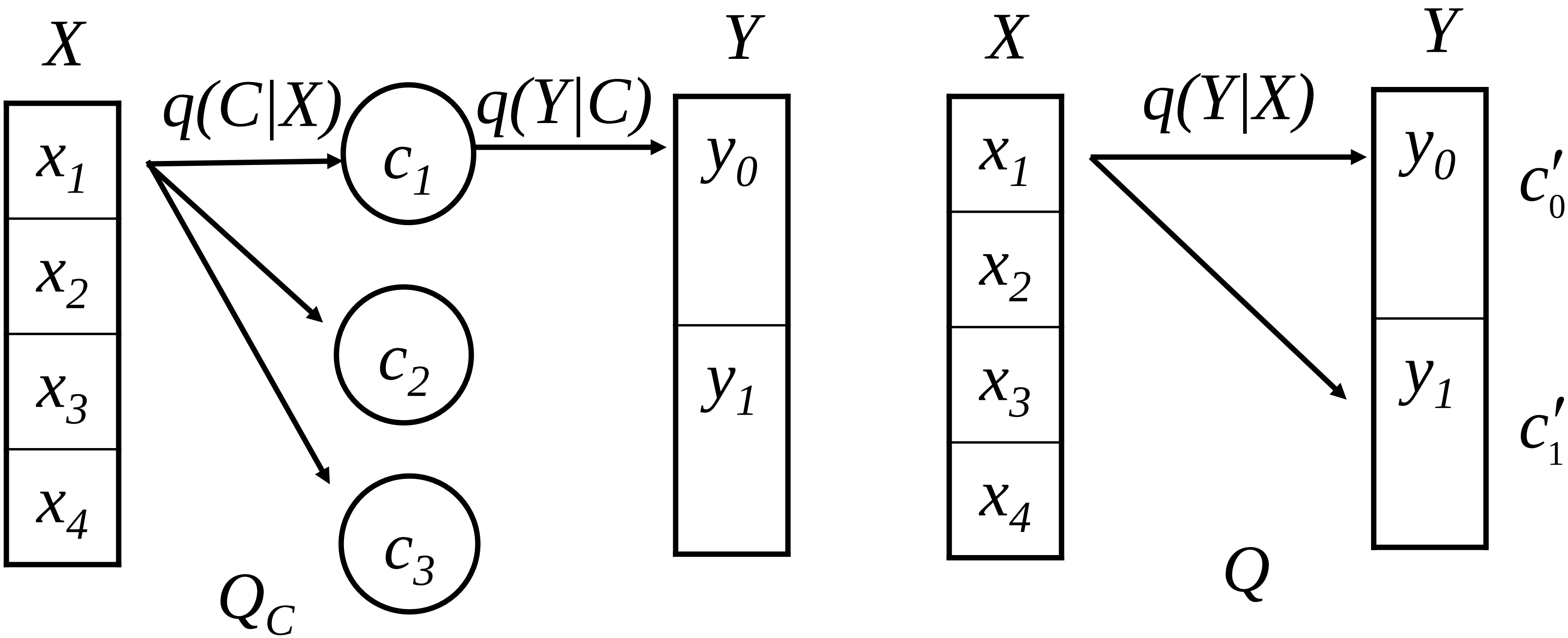
$$\text{Define}: q(y|x) = \sum_c q(c \mid x) q(y \mid c)$$

Optimality of Direct Mappings



$$\text{Define}: q(y|x) = \sum_c q(c \mid x) q(y \mid c)$$

Optimality of Direct Mappings



$$\text{Define}: q(y|x) = \sum_c q(c \mid x) q(y \mid c)$$

$$\hat{L}(Q_C) = \hat{L}(Q), \quad I_U(X;C) \geq I_U(X;Y)$$

A Bound for Direct Mappings

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{nI_U(X;Y) + K'}{2N}}$$

A Bound for Direct Mappings

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{nI_U(X;Y) + K'}{2N}}$$

- Tighter than the bound on $L(Q_C)$

A Bound for Direct Mappings

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{nI_U(X;Y) + K'}{2N}}$$

- Tighter than the bound on $L(Q_C)$
- Holds for any classification rule $q(Y|X)$

A Bound for Direct Mappings

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{nI_U(X;Y) + K'}{2N}}$$

- Tighter than the bound on $L(Q_C)$
- Holds for any classification rule $q(Y|X)$
- Can be optimized (gradient descent) with respect to $q(Y|X)$ to provide an optimal classification rule

A Bound for Direct Mappings

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{nI_U(X;Y) + K'}{2N}}$$

- Tighter than the bound on $L(Q_C)$
- Holds for any classification rule $q(Y|X)$
- Can be optimized (gradient descent) with respect to $q(Y|X)$ to provide an optimal classification rule
- No need for intermediate clustering
  - Only for a single feature

Some Insights

$$L(Q) \le \hat{L}(Q) + \sqrt{\frac{nI_U(X;Y) + K'}{2N}}$$

- $\hat{L}(Q)$ is minimized by $q_{ml}(y|x)$

Some Insights

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{nI_U(X;Y) + K'}{2N}}$$

- $\hat{L}(Q)$ is minimized by $q_{ml}(y|x)$
- For $I_U(X;Y)=0$, $L(Q)$ is minimized by $q_{ml}(y)$

Some Insights

$$L(Q) \leq \hat{L}(Q) + \sqrt{\frac{nI_U(X;Y) + K'}{2N}}$$

- $\hat{L}(Q)$ is minimized by $q_{ml}(y|x)$
- For $I_U(X;Y)=0$, $L(Q)$ is minimized by $q_{ml}(y)$
- Thus $L(Q)$ is minimized by smoothing $q_{ml}(y|x)$ toward $q_{ml}(y)$

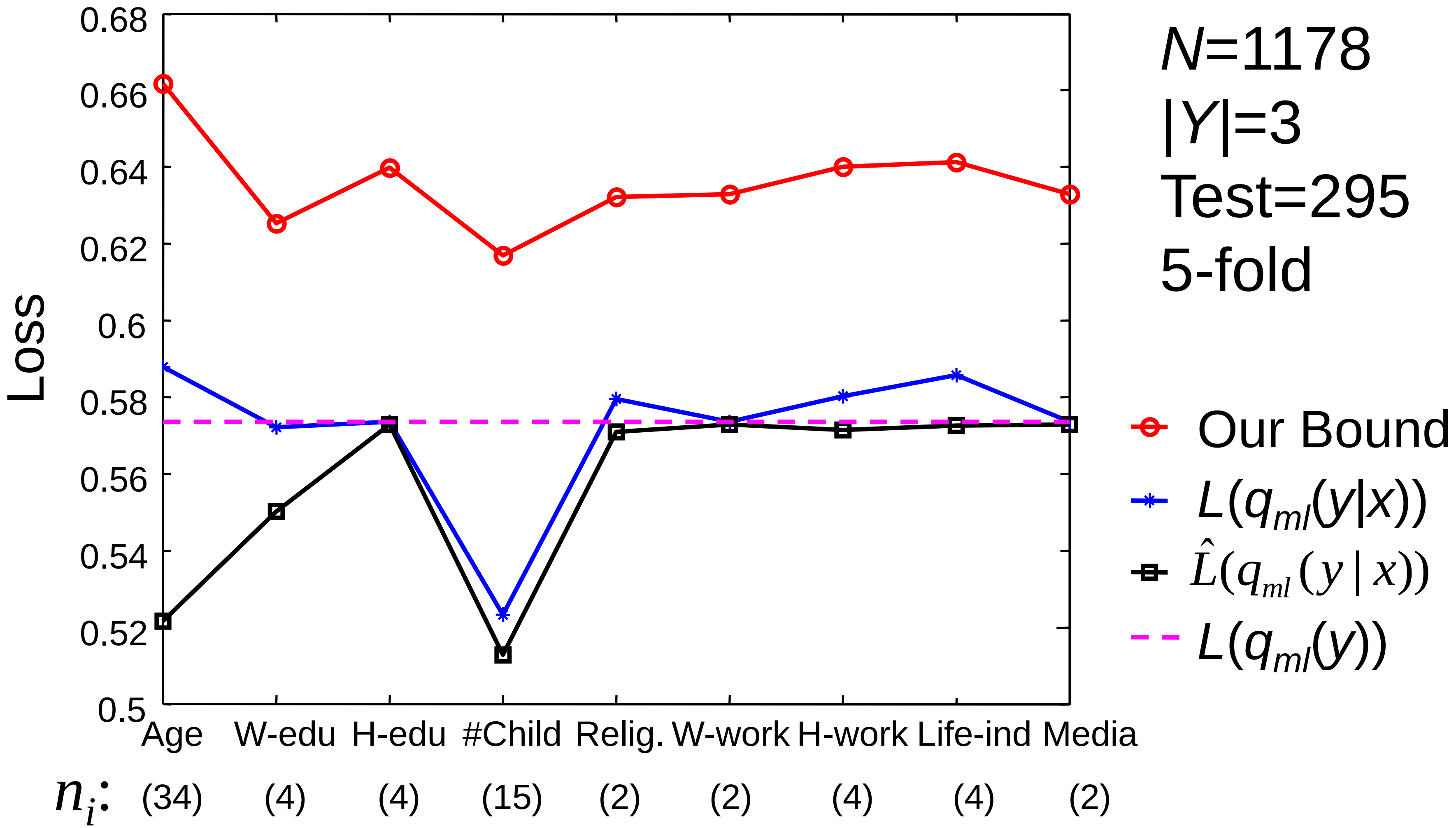
Application: Feature Ranking

- Rank features by their generalization potential
  - And not mutual information or correlation with the label
- Especially important for features of different cardinalities and small sample
  - Example: Y=cancer/no_cancer
    X_1=smoking/not_smoking
    X_2=year_of_birth

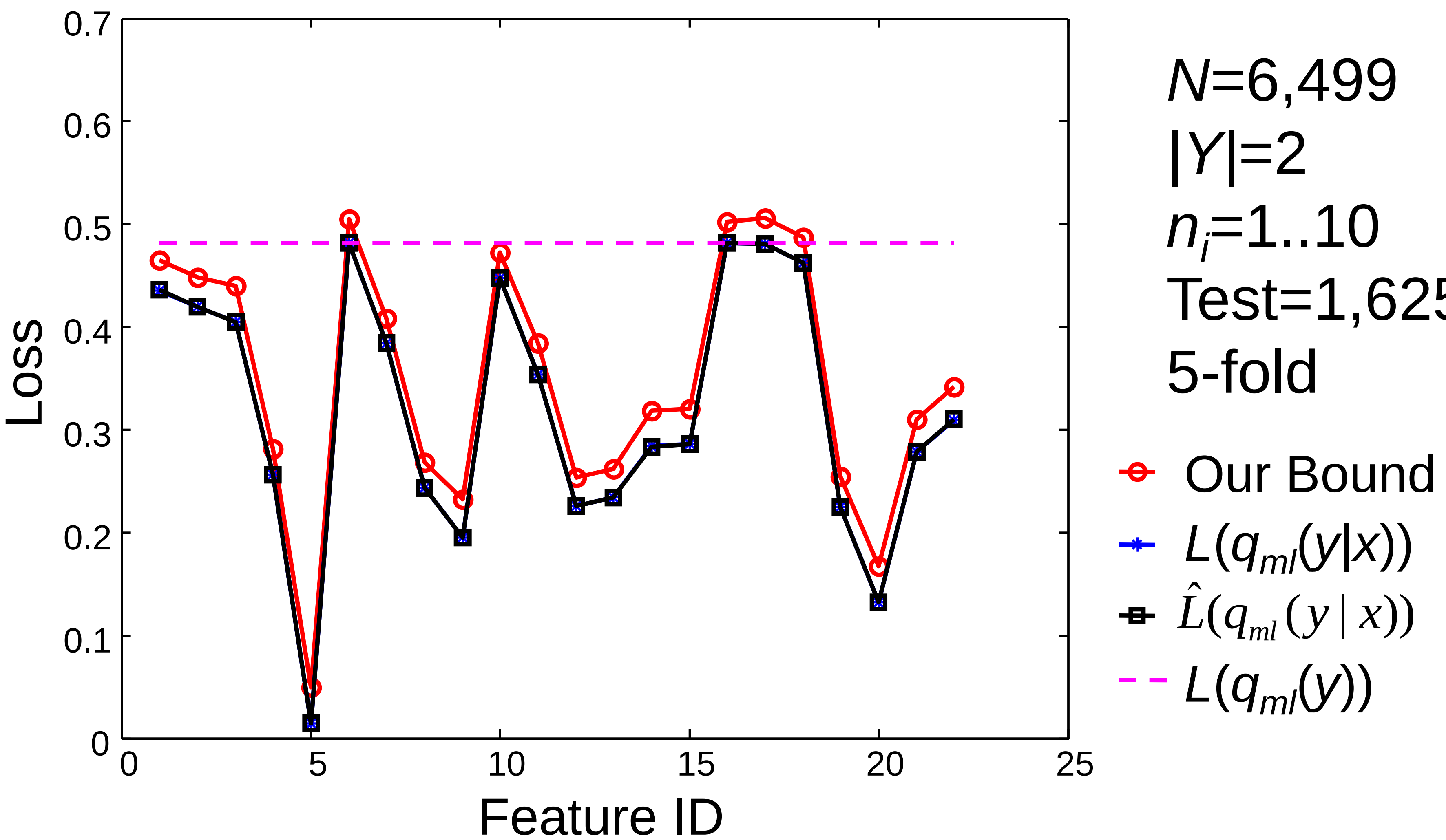
Related work

- Sabato, Shalev-Shwartz, COLT07
  - Generalization bound for $q(Y|X)=\hat{p}(Y|X)$ (empirical distribution)
- Our work:
  - Any $q(Y|X)$, in particular $q_{ml}(Y|X)$

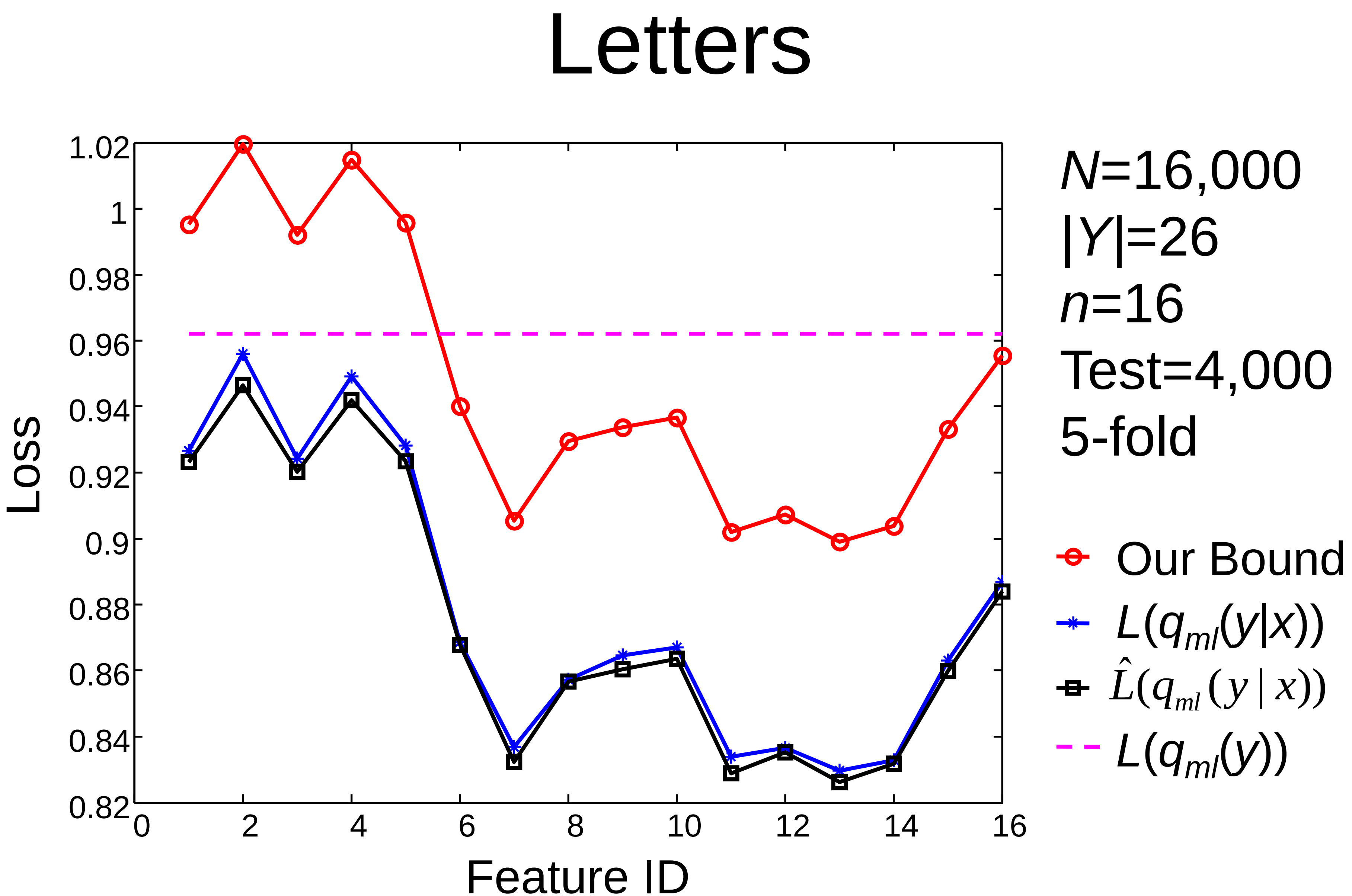
Contraceptive Method Choice



Mushrooms

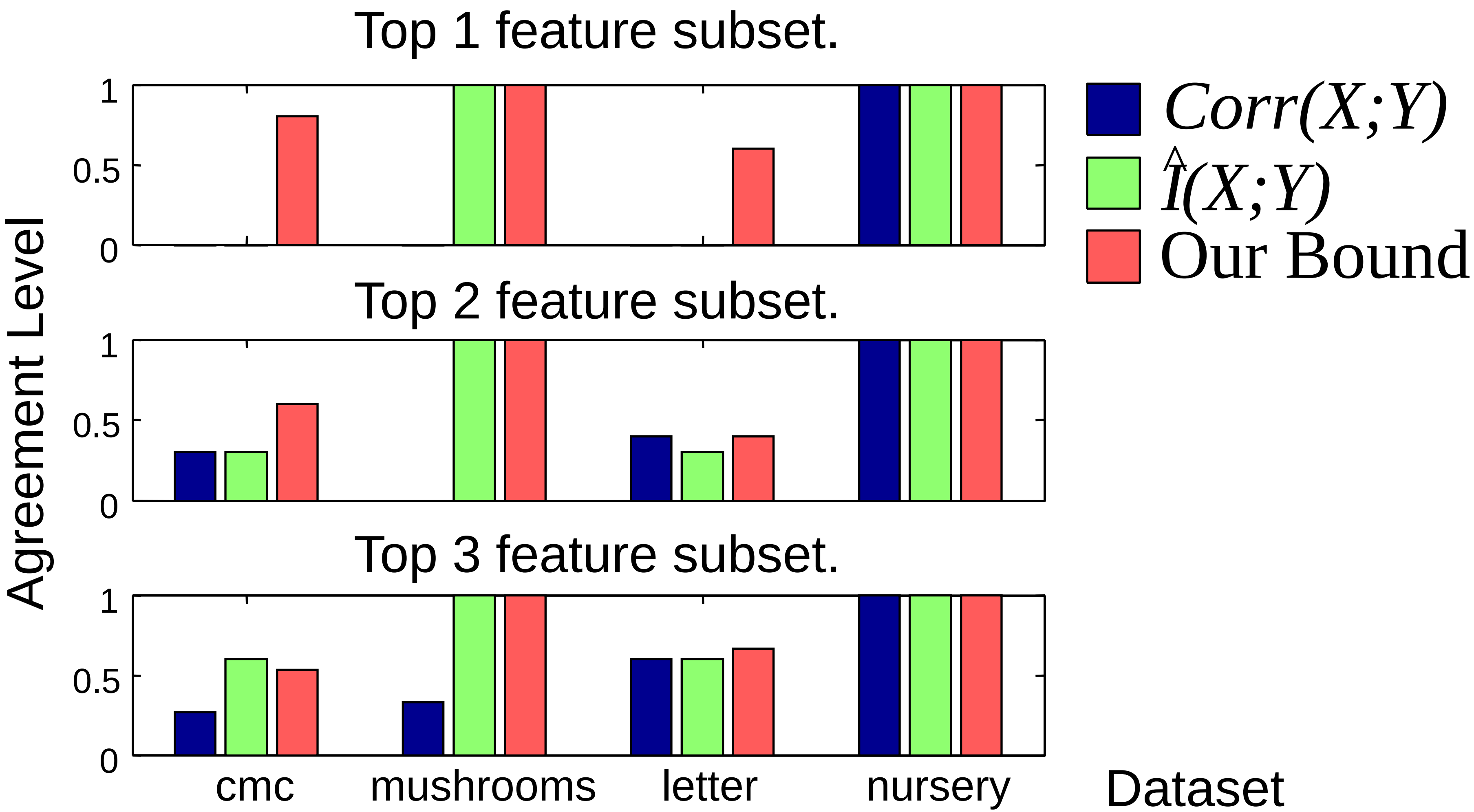


N=6,499
$|Y|$=2
n_i=1..10
Test=1,625
5-fold



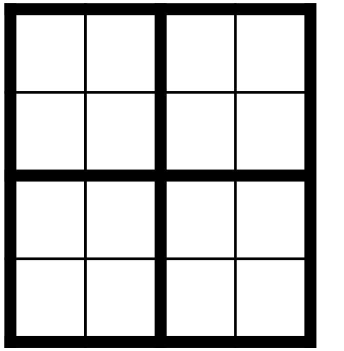
Letters
Loss
Feature ID
1.02
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0
2
4
6
8
10
12
14
16
N=16,000
|Y|=26
n=16
Test=4,000
5-fold
Our Bound
L(q_ml(y|x))
L̂(q_ml(y | x))
L(q_ml(y))

Comparison with MI and Corr



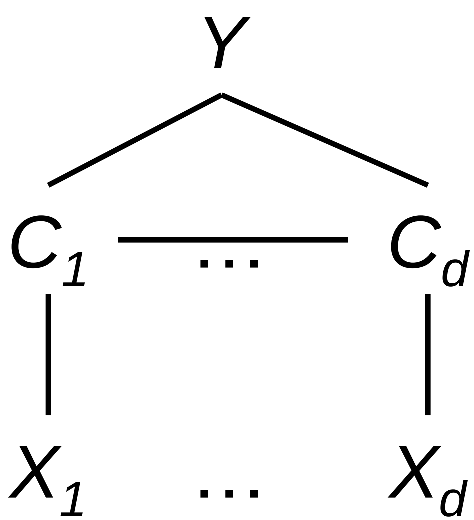
Summary

- Generalization bound for multi-classification based on grid clustering
- Unify feature values to amplify statistics
- Evaluation of clustering by its generalization power on a given task
- Feature Ranking
- Limitation: high dimensions

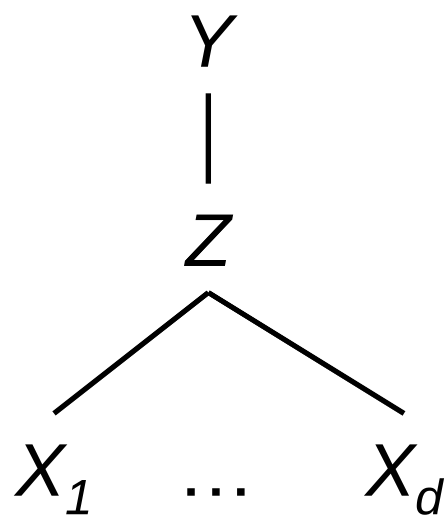


Future Work

- Derive a generalization bound for general graphical models



Grid Clustering



Factor Model